AFRL-HE-AZ-TP-2002-0006

UNITED STATES AIR FORCE RESEARCH LABORATORY

LARGE-SCALE LABORATORY TEST OF OCCUPATIONAL SURVEY SOFTWARE AND SCALING PROCEDURES

Walter G. Albert
Air Force Research Laboratory
Human Resources Directorate
2485 Gillingham Drive
Brooks AFB, TX 78235-5115

William J. Phalen
San Antonio, TX

David M. Selander
Air Force Research Laboratory
Human Resources Directorate
2485 Gillingham Drive
Brooks AFB, TX 78235-5115

Martin J. Dittmar
David L. Tucker
Daryl K. Hand
Johnny J. Weissmuller
Metrica, Inc.

Ian F. Rouse
Royal Australian Air Force

September 2002

20030917 117

Approved for public release; distribution is unlimited.

AIR FORCE MATERIEL COMMAND
AIR FORCE RESEARCH LABORATORY
Human Effectiveness Directorate
Warfighter Training Research Division
6030 South Kent Street
Mesa AZ 85212-6061

NOTICES

Publication of this paper does not constitute approval or disapproval of the ideas or findings. It is published in the interest of STINFO exchange.

Using Government drawings, specifications, or other data included in this document for any purpose other than Government-related procurement does not in any way obligate the US Government. The fact that the Government formulated or supplied the drawings, specifications, or other data, does not license the holder or any other person or corporation, or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

The Office of Public Affairs has reviewed this paper, and is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

WINSTON BENNETT JR.
Project Scientist

DEE H. ANDREWS
Technical Advisor

CURTIS J. PAPKE, Colonel, USAF
Chief, Warfighter Training Research Division

Direct requests for copies of this report to:

Defense Technical Information Center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, VA 22060-6218
http:// stinet.dtic.mil

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
September 2002

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*
Aug 1997 to Aug 1998

4. TITLE AND SUBTITLE
Large-Scale Laboratory Test of Occupational Survey Software and Scaling Procedures

5a. CONTRACT NUMBER
F33615-91-D-0010

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62205F

5d. PROJECT NUMBER
1123

5e. TASK NUMBER
A2

5f. WORK UNIT NUMBER
15

6. AUTHOR(S)
Walter G. Albert; William J. Phalen; David M. Selander; Martin J. Dittmar; David L. Tucker; Daryl K. Hand; Johnny J. Weissmuller; Ian F. Rouse

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Human Resources Directorate
2485 Gillingham Drive
Brooks AFB, TX 78235-5115

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Training Research Division
6030 South Kent Street
Mesa AZ 85212-6061

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL; AFRL/HEAS

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-AZ-TP-2002-0006

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Air Force Research Laboratory Technical Monitor: Dr. Winston Bennett, Jr., (480) 988-6561 Ext. 297, DSN: 474-6297

14. ABSTRACT
This paper describes a research and development effort to automate and improve data collection associated with USAF occupational surveys. Specifically the effort involves: (a) research and development of a PC-based procedure for self-administration of occupational surveys as a replacement for the existing paper-and-pencil process, (b) research and development of automated scaling procedures of optimal validity and reliability for obtaining measures of time spent on job-related tasks, (c) incorporation of feedback and branching techniques into the automated survey technology that will permit administration of large and complex occupational surveys, and (d) development of implementation guidelines for use in base-level computer systems and AF-wide electronic data transmission networks.

15. SUBJECT TERMS
Automated scaling; Computer-Adaptive survey; Data collection; Experimental scales; Job analysis; Occupational data; Occupational surveys; Scaling procedures; Software procedures; Task analysis

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON / 19b. TELEPHONE NUMBER *(include area code)*
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED	10	Ms Liz Casey 480.988.6561 x-188 DSN 474-6188

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Large-Scale Laboratory Test of Occupational Survey Software and Scaling Procedures

Walter G. Albert
Human Resources Directorate, Armstrong Laboratory

William J. Phalen
San Antonio, TX

David M. Selander
Human Resources Directorate, Armstrong Laboratory

Martin J. Dittmar, David L. Tucker, Daryl K. Hand, & Johnny J. Weissmuller
Metrica, Inc.

Ian F. Rouse
Royal Australian Air Force

INTRODUCTION

This paper describes a research and development effort to automate and improve data collection associated with USAF occupational surveys. Specifically, the effort involves: (a) research and development of a PC-based procedure for self-administration of occupational surveys as a replacement for the existing paper-and-pencil process, (b) research and development of automated scaling procedures of optimal validity and reliability for obtaining measures of time spent on job-related tasks, (c) incorporation of feedback and branching techniques into the automated survey technology that will permit administration of large and complex occupational surveys, and (d) development of implementation guidelines for use in base-level computer systems and AF-wide electronic data transmission networks.

The automated survey technology promises to provide higher quality data more rapidly for addressing urgent manpower, personnel, and training needs. Current methods of obtaining and processing data for occupational analysis are slow, complicated, and expensive, each step involving potential problems that can decrease sample size, lengthen projects, or introduce possible errors into a database used for Air Force decision making. This research should result in improved occupational data that will enhance management of Air Force specialties, estimates of training requirements, determination of the content of training courses, promotion selection, and job structuring. The accuracy of this information is becoming more crucial with the projected downsizing of the Air Force and the broadening range of responsibilities associated with various jobs.

A laboratory test of the software involving 572 randomly sampled subjects from 67 Air Force specialties (AFSs) has been completed. The steps in the computer-administered survey process are documented in Albert et al. (1993). The test was conducted at the Armstrong Laboratory's Experimental Testing Facility at Lackland AFB, TX. This facility provided a high degree of experimental control. Trained proctors administered the survey on forty identical PCs. The sample was selected such that higher and lower ability airmen were adequately sampled from technical and nontechnical specialties. For this study, lower aptitude was defined as having an Armed Forces Qualification Test score of 49 or less, and higher aptitude was defined as having a score greater than 49. This score separated the lowest quartile of airmen aptitudes from the upper three quartiles. Technical AFSs were defined as those having an Electronic or Mechanical score cutoff and nontechnical AFSs were defined as those having an Administrative score cutoff. AFSs having a General score cutoff were classified as technical or nontechnical depending on the degree to which the duties and tasks were technically or nontechnically oriented. To measure reliability of time spent estimation using an absolute time scale and four experimental scales, each job incumbent was administered the survey twice approximately two weeks apart. The remainder of this paper will discuss the criterion and experimental scales, results of the laboratory test, and future research.

EXPERIMENTAL SCALES

Research was conducted to determine the accuracy of four types of scales for obtaining job incumbent estimates of time spent on each task performed. The four scales were a Three-Stage Relative Time Spent Scale, a Direct Magnitude Estimation Scale, an Indirect Magnitude Estimation Scale, and an End-Anchored Graphical Scale. Time spent estimates from these four scales were compared with the criterion values, which were absolute time spent estimates based on a cross-product of two component measures: an estimate of absolute frequency of task performance and an estimate of the absolute amount of time normally required to perform the task once.

The first stage of the Three-Stage Scale is the same as the currently used nine-point, relative time spent scale. Each respondent was asked to estimate the relative amount of time he/she spends performing each task on a nine-point scale ranging from "very small amount of time" to "very large amount of time" compared to all other tasks he/she performs. At the second stage, the respondents were provided feedback in terms of groups of tasks to which they gave the same rating, so that they might refine the task ratings by using the groups of similarly rated tasks as contextual reference points for locating misrated tasks and moving them to task groups with more compatible time spent ratings. At the third stage, the refined groups of tasks were fed back to the respondent for further subdivision of each group of tasks into two or three more homogeneous subgroups, so as to yield up to 9 X 3 = 27 rating categories containing one or more tasks. The absolute time spent values for several high and low time consuming tasks were used to rescale the relative time spent values to absolute time spent values at each stage of the Three-Stage Scale.

The Direct Magnitude Estimation Scale required a mid-range task as an anchor point against which all other tasks to be scaled were compared by numerically estimating their time spent values as ratios of the anchor task's time-spent value. Several mid-range tasks from the absolute time set were used to rescale ratio estimates to absolute time.

For the first stage of the Indirect Magnitude Estimation Scale, the respondents used verbal anchors whose numerical values were previously derived by Direct Magnitude Estimation. At the second stage of this scale, the respondents located and moved misrated tasks as they did for the second stage of the Three-Stage Scale. Rescaling of estimates to absolute time for this scale was the same as for the Direct Magnitude Estimation Scale.

For the End-Anchored Graphical Scale, which required anchor tasks at both ends of the scale, the respondents rated each task by indicating its time spent value as a point on a horizontal line joining the two anchor tasks. The anchor tasks were the tasks having the highest and lowest absolute time spent estimates. Several additional tasks from the absolute time spent set were used to rescale this scale's ratings to absolute time spent values.

Subjects within each aptitude group and technical AFS vs. nontechnical AFS classification were randomly assigned to each experimental scale. For each scale, the job incumbent provided a total time spent rating on each task, each time spent rating was transformed to an absolute time spent estimate, and the job incumbent reviewed and made revisions to the absolute time spent estimates ordered high to low on time spent. The accuracy of each scale (and each stage of the Three-Stage and Indirect Magnitude Scales) was determined by comparing the rescaled time spent estimates of the experimental scale for each job incumbent with his/her edited absolute time spent estimates. For the Three-Stage Scale, equal-interval and ratio-interval estimates of absolute time spent were computed and presented for evaluation as two separate vectors.

RELIABILITY ANALYSES OF SCALES

CRITERION SCALE RELIABILITY

In order to produce a criterion value that would be comparable across scales, all reliability analyses were conducted, not at the task level, but at the case level, with the criterion for each case being the Fisher Z value corresponding to the correlation of each case's absolute time spent ratings across the two administrations. Fisher Z values were averaged for all cases in each scale type and reconverted to an average correlation ($\bar{r}$). Table 1 shows, for each experimental scale (treatment) and across all treatments, the number of respondents (N), the range, mean (M), and standard deviation (SD) of the number of tasks selected, the mean Fisher Z, $\bar{r}$, and the standard deviation of the Fisher Zs (SD_z). To get an acceptable measure of the reliability and validity of the time spent responses, a minimum of seven tasks was required to be selected by each subject; consequently, the data for eight subjects was excluded

from analysis because they responded to six or fewer tasks. The correlations of the absolute time estimates ranged from -.28 to 1.00. There were no significant differences ($p > .05$) among the mean Fisher Zs across scales.

Table 1. Absolute Time Spent Reliablity

Treatment	N	M/SD	Range	$\bar{Z}$ / SDz / $\bar{r}$
Three Stage	145	85/58	7-303	.74/.52/.63
Direct Magnitude	130	91/76	8-449	.86/.56/.70
Indirect Magnitude	137	85/62	7-366	.76/.63/.64
End Anchored	152	85/71	7-334	.79/.56/.66
Average	141	87/67	7-449	.78/.57/.65

In addition to the reliability measure shown in Table 1 (hereafter referred to as Z_A to denote that it was computed using information from all tasks selected by each respondent), five other measures of criterion reliability were computed: Z_W (Fisher Z for the weekly tasks), Z_E (Fisher Z for the essential tasks), APE_A (average proportional error for all tasks selected by the respondent), APE_W (average proportional error for the weekly tasks), and APE_E (average proportional error for the essential tasks). The average proportional error for subject j for any set of i tasks is defined as:

$$APE_j = \sum_{i=1}^{N_j} APE_{ij} / N_j$$

where N_j is the number of tasks responded to by case j and APE_{ij} is the absolute value of the difference in absolute time estimates at time 1 and time 2 for task i divided by the larger of the estimates.

Using Z_W, APE_W, Z_E, and APE_E as criteria, regressions were computed to determine if the reliability of absolute time estimates for the weekly and essential task subsets varied according to type of scale assigned to the subject. As expected, no significant differences were observed ($p > .05$). In addition, similar results were obtained with APE_A as the criterion as were obtained for Z_A. Finally, regressions were computed to see if the reliability of absolute time estimates for the subset of experimentally scaled tasks (up to 36 per rater) varied according to experimental scale assigned. Again, no significant differences were observed ($p > .05$). From these results, we can infer that the validity results presented later will not be affected by differential reliability being present among the sets of subjects assigned to each experimental scale.

EXPERIMENTAL SCALE RELIABILITY

For the Three-Stage Scale, two measures of reliability (Fisher Z and APE) were computed for each of the three stages and three data types (raw, interval scaled hours, and ratio scaled hours). There were no significant differences ($p > .05$) among the reliabilities at each stage. The mean Fisher Z varied from .59 (Stage 2, raw) to .70 (Stage 1, ratio) and the mean APE ranged from .24 (Stages 1 and 2, raw) to .66 (Stage 2, ratio). Transformation of the mean Fisher Z yielded a range of correlations from .53 to .60. The SDs of the Fisher Zs were large, ranging from .29 (Stage 1, raw) to .73 (Stages 1 and 3, ratio), and, similarly, the SDs of the APEs were also large, ranging from .12 (Stages 1 and 2, raw) to .21 (Stage 1, interval). In addition, chi-square results showed that respondents perceived interval-scaled data to be more accurate than ratio-scaled data ($p < .05$), although from a practical standpoint the preference was slight (57 percent at each administration).

For the Direct Magnitude Scale, the reliablity results were very similar for both the raw data (mean Fisher Z = .53, APE = .50) and the raw data converted to hours (mean Fisher Z = .55, APE = .51).

For the End-Anchored Graphical Scale, the reliablity results based on the Fisher Z statistic were also very similar for both the raw data (.77) and the raw data converted to hours (.80); however, the average proportional error was substantially less for the raw data (.36) than for the converted data (.56).

For the Indirect Magnitude Scale, there were no significant differences ($p > .05$) between the reliabilities at each stage. The mean Fisher Z varied from .87 (Stage 2) to .90 (Stage 1) and the mean APE ranged from .32 (Stages 1 and 2, raw) to .44 (Stages 1 and 2, hours). Transformation of the mean Fisher Z yielded a range of correlations from .70 to .72. The SDs of the Fisher Zs ranged from .36 to .39 and the SDs of the APEs ranged from .09 to .13. Therefore, across all scale/stage/data type combinations, Stage 1 of the Three Stage Scale gave the best results with APE as the measure of reliability. On the other hand, the Indirect Magnitude Scale gave the best results with correlation as the measure of reliability; however, the reliability associated with the End-Anchored Scale was not significantly lower ($p > .05$) than the reliability associated with the Indirect Magnitude Scale.

SCALE VALIDITY ANALYSES

VALIDATION OF CRITERION SCALE

Before discussing the validity of the experimental scales, the validity of the absolute time spent estimation procedure (criterion scale) must be established. First, it can be argued that the absolute time spent scale possesses superior content and construct validity relative to the four experimental scales. While all the experimental scales are focused on total time spent on a task, the absolute time spent scale decomposes total time into its two basic components: "time to perform a task once" and "frequency of performance." "Frequency" has been shown time and again in the literature to be a measure which corresponds to an innate counter mechanism we all possess, a perceptive ability that is consistently more accurate than that which governs our perception of time. In addition, "time to perform a task once" has the advantage of representing an average or median value rather than a total. For example, if you were to look at a long column of numbers, you could more quickly approximate a reasonably accurate median value than a sum for that set of numbers. Second, the absolute time scale permits the rater to respond unambiguously without need of translation or transformation of arbitrary scale values, such as those found in a 1-to-9-point relative time spent scale. Third, a task need not be compared with another task or "all other tasks" in order to make a frequency or time estimate. In other words, the absolute scale does not require task comparisons. Fourth, since the absolute time scale is not relativistic, neither is it ipsative, as is the 9-point relative time spent scale, which requires that each task be rated relative to "all other tasks I perform." Fifth, the absolute time scale sets no arbitrary limit on the magnitude of responses; whereas, the 9-point relative time spent scale severely limits the magnitude of a response, both in terms of the number of scale points available and the maximum weight a scale point can have as the number of tasks rated increases. In the final analysis, if we are willing to accept that the most valid scaling procedure is the one that allows the rater to say exactly what he/she wants to say with a minimum of ambiguity, then the absolute time spent scaling procedure would certainly merit the role of criterion as compared to the four experimental scales.

Two procedures were applied to provide empirical validation of the criterion scaling procedure. In the first procedure, every subject was presented pairs of time spent estimates in terms of hours per month for up to 10 tasks rated by that subject on both the absolute time spent scale and the assigned experimental scale. The tasks selected were those with the greatest discrepancy between the two estimates. The source of the estimates was not identified and the order of presentation was randomized. The subject was asked to select the estimate that he/she felt was the more accurate of the two. Preliminary analysis indicated that there was no bias toward selecting the first or the second estimate. If the criterion scale was truly more valid than any of the experimental scales, a significantly higher proportion of absolute time estimates should have been selected as more accurate than estimates derived from any of the four experimental scales. As expected, the absolute time spent estimates were selected more often than the experimental scale estimates, regardless of experimental scale, at both time 1 and time 2. The percentage of times that the absolute time spent estimate was chosen ranged from 52% (Direct Magnitude, time 1) to 56% (End Anchored and Three Stage, time 1). A value of $p < .001$ was associated with most of the computed chi-square values.

In the second procedure, each subject was presented the list of experimentally scaled tasks. The subject was asked to check those tasks he/she had performed within the last five working days. The results of this exercise were as follows: (1) At time 1, 70% of the tasks that had been checked as tasks performed at least once a week in the absolute time spent procedure were checked in this exercise as having been performed within the last five working days. At time 2, the percentage was 68%; a chi-square test across all cases (weekly vs. nonweekly and recently

performed vs. not recently performed) yielded a chi-square value with an associated probability of $p < .001$. (2) The correlation between tasks identified as performed or not performed within the last five working days at time 1 correlated .50 with time 2 identifications. (3) A t-test was computed between the mean number of times per year tasks were performed if they were identified as recently performed vs. the mean number of times per year for tasks not recently performed. Both time 1 and time 2 data yielded t-values with $p < .001$. The results of this procedure present evidence confirming the validity of the frequency measure used as a component of the absolute time spent scale and thereby provide a partial empirical validation of the criterion scaling procedure.

COMPARATIVE VALIDITY OF EXPERIMENTAL SCALES

The first step in determining the comparative validity of the four experimental scales was to identify the most valid form of each scale to use in the comparison. For all scales, the "best" functional form relating it to the criterion scale was sought. The forms considered were, $Y = a + bX$, $Y = a + b \ln X$, $Y = a + b_1X + b_2X^2$, $\ln Y = a + bX$, $\ln Y = a + b \ln X$, and $\ln Y = a + b_1X + b_2X^2$. For the Three-Stage Relative Time Spent Scale, additional alternatives had to be considered, such as which of the three stages was most valid and did ratio-interval or equal-interval rescaling of the ratings into estimated hours per year provide a better fit of the data to the criterion values. For the Indirect Magnitude Estimation Scale, a determination as to which of the two stages was most valid had to be made.

As for the reliability analyses, the validity analyses were conducted at the case level with the criterion value for each case being the Fisher Z value corresponding to the correlation of that case's experimental scale task ratings with the corresponding absolute time spent scale estimates. Fisher Z values were averaged for all cases in each scale type and reconverted to an average correlation. Time 1 and time 2 data were considered separately and combined. The results are shown in Table 2.

Table 2. Most Valid Form of Each Experimental Scale at Time 1 and Time 2 (For all scales, best functional form = parabolic)

	Mean correlation		
3-STAGE SCALE (N = 145)	TIME 1	TIME 2	TIME 1/TIME 2
STAGE 3 (RATIO INTERVAL)	.59	.62	.60
DIRECT MAGNITUDE (N = 130)	.58	.65	.62
INDIRECT MAGNITUDE (N = 137)			
STAGE 1	.66	.71	.69
END-ANCHORED (N - 152)	.85	.88	.87

Table 2 shows that for all scales, the parabolic equation ($Y = a + b_1X + b_2X^2$) provided the best fit of the experimental scale data to the criterion data; that the best fit for the Three-Stage Scale was the ratio transformation of data at Stage 3; and that the Stage 1 data of the Indirect Magnitude Scale provided a better fit than the Stage 2 data. It can also be ascertained from Table 2 that the End-Anchored Graphical scale provided a significantly better fit of the criterion than any of the other experimental scales. A t-value of 4.03 ($p < .001$) was computed for the difference between the combined time 1/time 2 correlation of .87 for the End-Anchored Graphical Scale vs. .69 for the Indirect Magnitude Scale (second best scale). The average correlation for the Indirect Magnitude Scale, however, was not found to be significantly different from the average correlations for the two remaining scales.

The second step in determining the comparative validity of the four experimental scales was to confirm the superior validity of the End-Anchored Graphical Scale by checking to see whether sampling biases regarding the types of jobs and job incumbents represented in the various experimental scale groups may have accounted for the differences in validity among the scales.

Although it was found that some variables, such as AFQT score and average time it takes the rater to complete the absolute time spent portion of the survey were important contributors to validity, in general, there was little criterion variance accounted for by these "nuisance" variables, and, after allowing for the variance accounted for by the job classification and job incumbent variables, the End-Anchored Graphical Scale was still found to be significantly superior to the other scales. This analysis did, however, find the Indirect Magnitude Scale to be significantly more valid than the remaining scales as a result of holding the job classification and job incumbent variables constant.

The major conclusions to be derived from these scale validity analyses are: (1) The absolute time spent criterion scale appears to be an acceptable criterion relative to the four experimental scales. (2) The End-Anchored Graphical Scale appears to be the most valid of the experimental scales. (3) The Direct Magnitude Scale and the nine-point relative time spent scale (Stage 1 of the Three-Stage Scale) appear to be the least valid of all the scale alternatives.

An appealing compromise scaling procedure that would use the End-Anchored Graphical Scale as the primary measurement device would be to have the rater employ the absolute time spent scaling procedure on a small subset of end-anchored-rated tasks covering the full range of time spent, thus enabling the conversion of the End-Anchored Graphical Scale ratings to estimates of absolute time by application of the parabolic functional relationship. Future R&D is planned to validate and refine this procedure as an operational spin-off of the computer-administered survey (CAS) system.

SUMMARY

This effort has produced a user-friendly, PC-based procedure for administering occupational surveys to job incumbents. By replacing the current hard-copy administration procedure, the time and cost (printing, mailing out, return mail, data entry) required to conduct occupational analyses will be greatly reduced, and a more effective use of resources such as manpower, time, and equipment will be possible. Quicker, more efficient turnaround time will meet Air Force managers' requirements for fast, accurate information on which to base critical manpower, personnel, and training decisions. In addition, the accuracy of individual and group job descriptions may be increased by adapting the data gathering process to the "intelligent" interactive, survey tailoring capabilities of a PC-based procedure and by the use of the most valid and reliable PC-based scaling procedures identified by the analyses reported in this paper. A questionnaire to assess the attitudes of raters toward the computer-administered survey (CAS) procedure yielded positive ratings concerning usage of the software.

REFERENCES

Albert, W. G., Phalen, W. J., Selander, D. M., Yadrick, R. M., Rouse, I. F., Weissmuller, J. J., Dittmar, M. J., & Tucker, D. L. (1993). Development and Test of Computer-Administered Survey Software. *Proceedings of the 35th Annual Conference of the Military Testing Association*, Williamsburg, PA.